ESTABLISHING THE VALIDITY OF SHORT-FORM COMPOSITE ITEMS IN THE CONTEXT OF TEACHING EVALUATIONS

BY

Gita Murli Sawalani

Submitted to the graduate degree program in Psychology
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

_____
Chairperson

Committee Members _____

_____

_____

_____

Date defended: August 6, 2008

The Dissertation Committee for Gita Murli Sawalani certifies that this is the approved version of the following dissertation:

ESTABLISHING THE VALIDITY OF

SHORT-FORM COMPOSITE ITEMS

IN THE CONTEXT OF TEACHING EVALUATIONS

_____
Chairperson

Committee Members _____

_____

_____

_____

Date approved: August 6, 2008

Abstract

The current study seeks to demonstrate the validity of a short form teaching evaluation instrument that has been created to measure several dimensions of teaching effectiveness. The primary motivation for this study stems from some of the shortcomings that exist with the current teaching evaluation tool that is used in the majority of classrooms at the University of Kansas, including its length. The items of the new short form are composite in nature such that each item consists of two or three key adjectives describing a particular construct of interest. The main empirical question is how much information would be lost by using this short form compared to a long form with multiple items per construct. For each student, data were collected on both a long and a short form. A total of 1297 students from 51 classrooms participated. Results indicated that the short form was a valid measure, despite smaller magnitude correlations and factor loadings compared to the long form measure of the same constructs.

Acknowledgements

First and foremost, I would like to thank God for all His blessings. I would like to thank my graduate advisor, Dr. Todd Little, for all his support and guidance throughout my graduate career, for the many research opportunities, and for being the chairperson of my dissertation committee. I would like to thank Dr. Kristopher Preacher for having faith in me, for being an excellent mentor in the short time that I have known him, and for being a member of my dissertation committee. I would like to thank Dr. Daniel Bernstein for all the wonderful conversations we have had regarding the scholarship of teaching and learning and for being a member of my dissertation committee. I would like to thank Dr. Patricia Hawley for broadening my horizons, making me think outside the box, and for being a member of my dissertation committee. I would like to thank Dr. William Skorupski for teaching me all I know about item response theory and for being a superb teacher. I would like to thank Dr. Neal Kingston for his methodological perspective on various topics and for being a member of my dissertation committee, despite the very short notice. I would like to thank my husband, Aaron Sauerwein, for his love and patience. Finally, I would like to thank my family and all my friends for believing in me. I could not have done this alone.

Establishing the validity of short-form composite items

in the context of teaching evaluations

Table of Contents

List of Tables

List of Figures

Establishing the validity of short-form composite items

in the context of teaching evaluations

The use of student evaluation instruments in the assessment of university courses and teaching effectiveness has become commonplace over the past few decades (Amin, 2002; Blackhart, Peruche, DeWall, & Joiner, 2006). This widespread practice of relying on student feedback is not limited to the United States (Byrne, 1992) but is adhered to in several Western European countries including the United Kingdom, The Netherlands, France, and Germany (Husbands & Fosh, 1993).

Despite the extensive use of student ratings in colleges and universities, there has been considerable debate regarding their utility. Some argue that they are useful since students are exposed to the instructor's teaching over the length of a course and that students are in fact the constituency that should be making such judgments. Others argue that students have an inherent conflict of interest in making such judgments and are influenced by lenient grading and entertainment value. As a result, students may not be appropriate judges of teaching practices that facilitate student learning (Brown, 1976). More recently, scholars of teaching and learning have argued that students' perspectives are important regardless of the inherent bias. In this context, both teaching and learning should be based on multiple forms of evidence including, but not limited to, student ratings. Two key questions that arise include: (a) what dimensions of teaching should students evaluate and (b) how should these ratings be best obtained? My dissertation project stems from my involvement in a

university project that was established to address these two questions. Before addressing these two questions, a little background information is needed.

Background: Task force on the assessment of teaching and learning

The task force on the assessment of teaching and learning was established in July 2006 by the Faculty Senate Executive Committee (FacEx) at the University of Kansas to consider the current evaluation process of teaching and learning and to propose guidelines to aid academic units in evaluating teaching and learning. The task force recommendations can be broken down into three main categories. The first recommendation involves faculty members reporting on a broad range of teaching-related activities, such as how they conduct or prepare for a course, as the basis for peer review. The second recommendation involves student ratings of teaching effectiveness that should be concise and focused on the aspects of teaching that students likely know best. The third and final recommendation involves open-ended comments to guide faculty members in the improvement of teaching. The focus of the current study will be on addressing the second recommendation (i.e., student ratings of teaching effectiveness).

Because the task force on the assessment of teaching and learning recommended that student ratings of teaching effectiveness should be concise and focused on the aspects of teaching that students likely know best, two tasks needed to be addressed: (a) what are the dimensions students should rate and (b) can a concise instrument be developed that can reliably and validly measure these dimensions?

Regarding the dimensions to include, the task force outlined a set of six key facets based on the Kansas Board of Regents mandate. Specifically, in terms of what students are likely to know best (or the knowledge that students are most likely to have), the task force on the assessment of teaching and learning concluded that students are good candidates to evaluate: (1) clarity and organization of classroom time, (2) faculty support and availability, (3) clarity and organization of course materials, (4) setting and meeting goals and expectations, (5) maintaining a respectful climate, and (6) perceptions of their own learning. These factors coincide with the minimum requirements put forth by the Kansas Board of Regents. More specifically, the Kansas Board of Regents states that:

> "Instruments to measure student ratings of instruction should solicit, at a minimum, student perspectives on (a) the delivery of instruction, (b) the assessment of learning, (c) the availability of the faculty, and (d) whether the goals and objectives of the course were met".

Shortcomings of the current teaching evaluation instrument

Some of the shortcomings of the current teaching evaluation instrument (Appendix A) include: (1) its length – the current form has a total of 46 items on three differently worded Likert-type scales, (2) questions that do not apply to all fields (e.g., art students might have projects instead of exams), (3) questions that might apply to undergraduates but not graduate students and vice versa, (4) questions that might present a conflict of interest (e.g., "The readings were too difficult"), (5) a lack of theoretical cohesiveness across the 46 items, and (6) two "overall" questions –

often over-weighted and misinterpreted in practice. Moreover, it is unclear whether the current form adequately addresses all the dimensions mandated by the Kansas Board of Regents because of the awkward and vague wordings of many items (see Appendix A). For example, under the evaluation section of the current form, one of the items reads "The objectives of the course and the methods are clearly explained". It is unclear, for example, what objectives the "methods" would have nor whether the "methods" are something to be clearly explained.

Goals of the current study

The goals of the current study are to develop a short teaching evaluation instrument that consists of short-form composite items and to establish the reliability and validity of these items. A key question is "How much damage would result from using a shorter form with a composite item for each key construct rather than a comprehensive long form?" In other words, "Can a short-form composite item capture a similar amount of information compared to having multiple items per construct?" This study tests this question empirically.

Single-item indicators

Single-item indicators can be classified into two distinct groups. The first group consists of single-item indicators that are designed as single-item measures of a particular construct. The second group consists of global single-item indicators that require participants to consider all aspects of a phenomenon and provide an overall rating of this phenomenon (Youngblut & Casper, 1993). Items in this category typically involve the word 'overall' and do not specify the specific aspects of a

phenomenon being rated. Items in the first category, on the other hand, typically provide a specific referent to the domain being measured and do not involve the word 'overall'. The short-form composite items in the current study are similar to single-item indicators outlined in the first group in that each item taps into a particular construct, references a specific feature of the construct to be rated, and do not include the word 'overall'. However, the current items extend this category somewhat by including more than one specific feature of the construct being rated. Specifically, each item includes 2 or 3 adjectives pertaining to the construct of interest. For example, for the instructor's teaching construct, the item reads, "The instructor's teaching was clear, understandable, and engaging".

The approach described in the second group of single-item indicators (e.g., 'overall' items) is precisely what the task force on the assessment of teaching and learning is trying to avoid. Hence, there was a desire to eliminate the two "overall"-rating items on the current teaching evaluation instrument. Two critical questions to address with a measurement approach that relies on single-item indicators include (a) are they reliable and (b) are they valid?

Reliability

Reliability refers to the accuracy or precision of measurement (Cronbach, 1951; Widaman, Little, Preacher, & Sawalani, 2008). Reliability is usually defined as the ratio of true score variance relative to total variance of a particular scale of interest (i.e., $\sigma^2_{\text{true score}}/\sigma^2_{\text{total}}$). As this ratio increases, error variance decreases and measurement become more accurate (Widaman et al., 2008). The more items there

are in a given scale, the more this ratio increases because the common variance among the items is aggregated leading to a more reliable the measure. That is, this increase in reliability is because as more items are added together, more true score variance (relative to error variance) is reflected in the sum score[1]. The variance of the sum of two items, for example, is equal to the sum of the two variances plus two times the covariance between the two items (i.e., amount of true score variance common to the two items; Hill & Lewicki, 2005).

Several different types of reliability measures exist. These include: (a) split-half, (b) internal consistency, (c) parallel forms, and (d) test-retest reliability. In terms of internal consistency, coefficient α is by far the most common internal consistency reliability measure used in the social sciences (Cronbach, 1951). Unfortunately, with only a single item and a single assessment occasion, such traditional measures of reliability become impossible to calculate.

An alternative way to consider reliability of a short-form composite item is to examine its strength of association with another similar measure or indicator of the construct of interest. Specifically, the correlation of a short-form composite item with a long form scale that assesses the same construct would provide a lower-bound estimate of the reliability of a short-form composite item. The basic idea here is that the degree of association with any other external measure reflects reliable information. The amount of this information would be an estimate of the minimum reliability of the short-form composite item. This estimate would be of the minimum reliability because the variance in the long form and the short form that is related to

the method of assessment (and thus not correlated with each other) would attenuate the true association between the two forms. Thus, the actual reliability of a short-form item is likely to be larger.

Another estimate of reliability of a short-form item is the magnitude of its loading in the context of a confirmatory factor analysis (CFA). As I describe in more detail below, a CFA model that has indicators of a common construct gathered from both the long and short form would allow one to control for some of the attenuating influence of non-shared method variance. Because a loading in a CFA model is the amount of variance in an indicator that is explained by the construct, the magnitude of the loading would reflect the amount of reliable variance of the short-form indicator that is associated with the construct of interest.

Validity

Validity involves whether or not the scale or measure does a good job at measuring the construct that we want to measure. There exist several different types of validity, including: (a) content, (b) criterion, and (c) construct validity. According to Messick (1989), validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". Traditionally, since the early 1950s, validity has been divided into three distinct types – content, criterion, and construct validities (Messick, 1989, 1995). In his 1995 paper, Messick describes a more comprehensive view of validity that encompasses considerations of content, criteria, and consequences in a construct

framework for the purposes of testing hypotheses. Within this framework, content and criterion validities are considered as part of construct validity (Messick, 1995).

Content validity refers to how well a scale includes content that relates to all aspects of the dimension being measured. For example, with regard to the instructor's teaching construct, the scale should contain different aspects of the instructor's teaching, such as clarity and organization. Criterion validity is determined by examining the associations of a particular scale of interest with key variables that are identified as criteria. The magnitude or pattern of such associations should be consistent with prior findings that exist in the literature (Widaman et al., 2008). Finally, construct validity refers to the amassed evidence that a measure is an accurate measure of the theoretical construct it is meant to measure. Here the consistency of the evidence from the content and criterion relationships as well as the performance of a measure across different assessment conditions contribute to the overall construct validity of a given construct. I will describe in more detail how I plan to examine each type of validity in the methods section.

Factors that influence student ratings

As a form of criterion validity, one of the key characteristics of the short-form composite items that I will evaluate is their behavior with regard to known factors that influence student ratings. That is, I will evaluate whether the short form is influenced in a similar way as other established measures. The criterion-validity factors that I will examine include instructor reputation, class attendance, and expected grade.

Instructor reputation

Of all the studies reviewed, only one specifically looked at the relationship between instructor reputation and student ratings. This study was a study conducted by Griffin (2001) which examined the relationship between instructor reputation (as perceived by their students) and student ratings of both the course and teaching effectiveness. Based on what the students reported about the instructor's reputation prior to enrolling in the course, a student was assigned to one of the following three groups: positive reputation, no information, and negative reputation.

Data were presented for nine education classes for an overall instructor rating and an overall course rating. For the overall instructor rating, the mean for the positive reputation group was the largest for all nine classes. Furthermore, significant differences for groups were found for the majority (i.e., six out of nine) of classes. The mean effect size for the positive reputation group was $d = 0.82$ and $d = -0.40$ for the negative reputation group. The effect size is calculated as follows: $d = [M_{\text{positive or negative}} - M_{\text{no information}}]/ SD_{\text{overall}}$. The positive reputation indicator correlated 0.22 with the overall instructor rating. The correlation was -0.36 for the negative reputation indicator.

In terms of the overall course rating, the mean for the positive reputation group was the largest for seven out of nine classes, with the no information group having the largest mean for two of the classes. Significant difference among the groups were found for three (out of nine) classes. The mean effect size for the positive reputation group was $d = 0.42$ and $d = -0.63$ for the negative reputation

group. The positive reputation indicator correlated 0.20 with the overall course rating, while the negative reputation indicator correlated -0.30.

Based on the fact that the means were generally the highest for the positive reputation group the majority of the time, I expect to see a significant positive relationship between instructor reputation and student ratings on the proposed short-form measure, particularly for the instructor-related dimensions (see instrument description below).

Class attendance

A few studies have also considered the impact that class attendance has on students' ratings of their instructors. A study conducted by Burns and Ludlow (2005) showed that students' perception of whether regular attendance is necessary was a significant predictor of positive instructor ratings. More specifically, according to their study, the relationship between students' perception of class attendance and instructor ratings accounted for 5.3% of the variance after controlling for other factors such as class size, instructor availability, and small-group interactions. Based on this finding, I expect to see a significant positive relationship between class attendance and student ratings, even though Burns & Ludlow (2005) focuses on students' perception of whether attendance is important and the current study focuses on actual attendance based on self-report. This expectation is based on the assumption that actual attendance would be a correlated proxy for the perception that attendance is important.

Expected grade

A third factor that has been found to influence student ratings is their expected grade. The majority of studies reviewed (Bausell & Magoon, 1972; Ginexi, 2003; Holmes, 1971; Krautmann & Sander, 1999; Maurer, 2006; Stumpf & Freedman, 1979) found a positive relationship between expected grade and course evaluations. Holmes (1971), for example, found a positive relationship between expected grade and the degree to which students were stimulated by the instructor and felt that the grading system was fair. Items assessing the instructor's presentation, however, were not found to be related to expected grades. Maurer (2006) found that students who expected to receive a D in the course rated their instructor lower than students who expected an A, B, or C. Students who expected a C rated their instructors lower than students who expected an A.

Only a few studies reviewed (Blum, 1936; Garverick & Carter, 1962; Marlin & Gaynor, 1989) found no relationship between expected grade and course evaluations. Marlin and Gaynor (1989), for example, found that overall evaluation of instructors by students seems to be based primarily on assessment of instructor teaching behaviors and not other variables including expected grade. Although they did not find a significant relationship between anticipated grade and the evaluation of the instructor, they found some relationship between the evaluation of the instructor and the anticipation of a grade lower than expected. Based on the fact that the majority of studies found a positive relationship between expected grade and student

ratings, I expect to see a significant positive relationship (in terms of a simple correlation) between the two variables.

In summary, my dissertation will address the question of the dimensions that students should be rating and whether or not the short-form measure will be able to measure these dimensions in a reliable and valid manner.

Method

Participants

Approval for this study was obtained through the University of Kansas (KU) Internal Review Board. A total of 1297 students participated. Data were collected from a convenience sample of 51 classes from a variety of departments at KU during the Spring and Summer semesters of 2007. These departments included the Departments of Psychology, Communication Studies, Psychology and Research in Education, Curriculum and Teaching, Health Sport and Exercise Sciences, Educational Leadership and Policy Studies, and Special Education. Table 1 depicts the breakdown of all the classes by department. Additional information for each class include the class number, the line number (i.e., a unique 5-digit number assigned to each course at KU), the semester and year, the number of students who participated in the study, and whether it was an undergraduate- or graduate-level course. The total sample size of all 51 classes does not add up to exactly 1297. This discrepancy is due to 8 students who did not fill in the line number for the course and 20 who incorrectly provided the line number. Though these 28 forms were included in the analyses, they are not reflected in Table 1.

Measures

Based on the minimum requirements recommended by the Kansas Board of Regents, six constructs were developed. These constructs included: (1) an instructor's teaching construct; (2) a learning construct; (3) a help construct; (4) a goals and objectives construct; (5) a materials and content construct; and (6) an expectations construct. For each construct, a large pool of items was created and several committee meetings at the Center for Teaching Excellence at KU were held to discuss and finalize the item list.

Given the six distinct dimensions that are required to be assessed, the current study followed a two-pronged approach. First, a long form (Appendix B), following traditional psychometric practices, was created to assess each of the six dimensions[2]. Second, in accordance with the goals of the task force, a short-form version (Appendix C) was developed. For this version, each of the six dimensions is represented by a sentence with multiple adjectives (e.g., "The instructor's teaching was clear, understandable, and engaging"). The use of multiple adjectives was intended to ensure full coverage of the key aspects of each focal dimension. The idea here is that students would combine the content covered by the adjectives in making their assessment, thereby providing a more comprehensive evaluation than if only one adjective was used. This approach attempts to combine the merits of multiple indicators of a construct with those of a single-item measure. This type of question format is contrasted with the more psychometrically sound measure of the same dimension (e.g., a 12-item scale where each item contains only one adjective).

In order to assess the psychometric properties of the short form, the data for my dissertation were collected for both the long and short forms from each student. This procedure allows examination of the content validity and reliability of the short form in comparison to the long form.

The questionnaire protocol included additional questions of interest such as the student's perception of the amount he or she has learned compared to other similar courses[3], the importance of specific reasons for taking a particular course (e.g., course fulfills a major or minor requirement), student status (i.e., undergraduate, graduate, other (faculty, staff, non-degree)), year of study, how often the student completed required coursework, how many times a week the class met, what grade the student expected to get in the class, and how many class periods a student had missed over the course of the semester (see Appendix C).

A secondary question in this project is whether using a 5-point versus a 7-point scale for the responses would yield different reliabilities in the scale scores. For the 5-point scale, the response options were: "strongly disagree", "disagree", "neither agree nor disagree", "agree", "strongly agree". For the 7-point scale, the options were: "very strongly disagree", "strongly disagree", "disagree", "neutral", "agree", "strongly agree", and "very strongly agree". Because student ratings are generally quite reliable, I expect the estimated reliabilities derived from responses on the 5-point scale to not differ meaningfully from those derived from the 7-point scale. This question is addressed by examining the internal consistency estimates of the indicators from the long form.

Counterbalancing

For the purposes of counterbalancing, 659 students took the long form first and the short form second (n=344 on a 5-point scale; n=315 on a 7-point scale) and 637 students took the short form first and the long form second (n=323 on a 5-point scale; n=315 on a 7-point scale). One reason for counterbalancing was concern that the order of presentation might influence the quality of the responses. For example, students might get bored when responding to the protocol if they completed the long form measure first. Conversely, if they get the short form first, they might not feel motivated to fill out the long form diligently. However, because the long and short form together are relatively quick and easy to complete, I do not expect any order effects between the short and the long forms. I will test for these order effects by examining mean levels and correlations across the counterbalanced protocols. Specifically, I will compare short-form responses that were obtained prior to long-form responses with those that were obtained following the long form. Likewise, I will compare long-form responses that were obtained prior to short-form responses and long-form responses that were obtained after short-form responses.

Data manipulation

All missing data were imputed using the SAS PROC MI procedure. All inferential statistical analyses were performed using the imputed data sets while the descriptive statistics are based on the unimputed data. For each long-form construct, one parcel was created using the items that appeared in the short-form composite item. The item-to-construct balance method proposed by Little, Cunningham, Shahar,

& Widaman (2002) was used for the remaining items for each construct resulting in at least three parcels per construct. For example, for the "teaching" construct, one of the parcels created included the following items: "clear", "understandable", and "engaging".

A parcel is a simple average of several items assessing the same construct (Kishton & Widaman, 1994). Parcels were created primarily due to the disadvantages associated with analyzing data at the item level. Item-level data tend to be less reliable and are more likely to violate distributional assumptions. Models based on parcels tend to be more parsimonious (both locally, in terms of defining a construct, and globally, in terms of representing the full model), have fewer chances for correlated residuals or for an item to load onto more than one construct, and lead to reductions in sampling error (Little et al., 2002).

Prior to creating parcels, a few items were dropped for a variety of reasons. The item "curt" of the help construct, for example, was dropped because quite a few students appeared to be unfamiliar with the word. This item yielded contradictory responses compared to other items that had a negative connotation. In other words, students were not sure if being curt is positive or negative. Similarly, the item "simplistic" of the content construct was dropped as students were not sure if being simplistic is positive or negative. Finally, the item "explicit" of the goals construct was dropped primarily due its low correlation with other positively worded items for this construct.

Before proceeding to the analyses, I first examined whether there were differences in the reliability estimates between the 5-point and the 7-point scales. As shown in Table 2, Cronbach's α values obtained for the long form on 5-point and 7-point scales were quite similar. Reliability coefficients for all constructs on both scales were in the 0.80 to 0.90 range (Table 2). As a result, the data collected on the 7-point scale were rescaled to be on a 5-point scale. Specifically, the 1 to 7 response scale was rescaled to a 5-point scale using the following formula:

$$r5 = \{[(r7 - 1)/6] * 4\} + 1 \quad (1)$$

where r5 is the resulting response rescaled to a 5-point equivalent and r7 is the observed response from the original 7-point scale.

Procedure

Content validity. To examine the content validity of the short form, I will look at the means and the standard deviations for the six constructs[4] of the short and long forms. I will also examine the correlations between the composite items of the short form and the parcels of interest of the long form (i.e., the parcels that contain the same items as the short-form composite items - one parcel for each construct). That is, because data were collected on both the long and short forms for each participant, I can examine the content validity of the short-form composite items in relationship to the multi-item measure (Youngblut & Casper, 1993). Besides means, standard deviations, and correlations, I will examine the content validity of the short form by considering a confirmatory factor analysis model (Figure 1).

Confirmatory factor analysis (CFA) is a form of structural equation modeling (SEM) that deals with the relationship between observed measures or indicators and latent variables or factors. CFA is quite different from the more familiar exploratory factor analysis (EFA) in that the researcher has to specify all aspects of the model *a priori* (Brown, 2006).

CFA has become one of the most popularly used statistical procedures in applied research. This popularity results from CFA being able to answer many different types of research questions. CFA is commonly used to examine the content validity of a set of *a priori* specified factors by examining their relationships with observed indicators. The CFA model can also be used to inform construct validation. Specifically, convergent validity would be supported when the various indicators of overlapping constructs are interrelated, and discriminant validity would be supported when the indicators of distinct constructs are not interrelated above and beyond the degree of association contained at the construct level. The CFA model is also useful for estimation of potential method effects (i.e., where covariation among indicators is not due to the factor, but rather due to the measurement approach utilized) and for examining measurement invariance (i.e., how well a particular measurement model generalizes across groups or time; Brown, 2006).

Besides considering a single-group CFA model that considers all 6 dimensions of teaching effectiveness simultaneously, I will perform two separate two-group CFAs comparing the following: (a) data collected on the 5- versus the 7-point scales and (b) data collected on the 5- versus the rescaled 5-point scales. For the

5- versus 7-point scales, I do not expect to find any differences in the factorial structure (i.e., the indicator-to-construct pattern for both factor loadings and intercepts) of the measurement model. However, I do expect to find differences in the variances, covariances, and means. For the 5- versus rescaled 5-point scales, I do not expect to find any differences in the factorial structure, variances, covariances, or means.

For each of the two-group CFA models (i.e., 5- versus 7-point scales and 5- versus rescaled 5-point scales), I tested the following steps in sequence: (a) a test of the initial configural model that specifies the relationship between manifest indicators (i.e., observed variables) and latent constructs (i.e., unobserved variables), (b) a test of the measurement equivalence in the measurement of these models across scales (specifically in terms of equating the loadings and intercepts of the observed variables across the two scales), (c) a test of the homogeneity of the variances and covariances of the latent constructs for the two scales, (d) a test of the homogeneity of variances only (if homogeneity of variances and covariances are not obtained), (e) a test of the equivalence of means of the latent construct for the two scales, and (f) a test of the equivalence of the patterns of correlations for the two scales (if homogeneity of variances and covariances are not obtained). For both two-group CFA models (i.e., 5- versus 7-point scales and 5- versus rescaled 5-point scales), I do not expect a difference in the magnitude of the correlations of the latent constructs because the relationship among latent constructs should remain unchanged regardless of the scale

on which data were collected. This test of equivalence of the patterns of correlations will be performed by creating phantom constructs for each of the latent constructs.

For the test of the measurement equivalence (in terms of equating the loadings and intercepts of the observed variables across the two scales), I will consider the change in the comparative fit index (CFI). The decision to use this particular goodness-of-fit index is based on the fact that the most common goodness-of-fit index (i.e., the χ^2 statistic) is dependent on (or too sensitive to) sample size and is a test of exact fit. The question of whether factorial invariance holds or not is generally a question of approximate model fit, not exact fit (Little, 1997). For large sample sizes (such as in the current study) the χ^2 statistic provides us with an overly sensitive statistical test of model fit, but not necessarily a practical test (in terms of the measurement model) of model fit (Cheung & Rensvold, 2002; Little, 1997).

A simulation study performed by Cheung and Rensvold (2002) proposed that a change in CFI smaller than or equal to 0.01 suggests that the null hypothesis of invariance should not be rejected indicating that there are no differences between the two scales (in terms of loadings and intercepts of the observed variables; Cheung & Rensvold, 2002). More recently, Meade, Johnson, and Braddy (2008) recommended a more lenient cutoff of 0.02. For the purposes of the current study, I will use the cutoff of 0.02 as recommended by Meade et al. (2008).

For the test of the homogeneity of the variances and covariances of the latent constructs for the two scales, I will consider the difference in χ^2 between the intercept invariant model (if constraints are found to be tenable) and the model in which all

variances and covariances of latent constructs are equated across the two scales. For the test of homogeneity of variances only (if homogeneity of variances and covariances are not obtained), the test of the equivalence of means of the latent construct for the two scales, and the test of the equivalence of the patterns of correlations for the two scales (if homogeneity of variances and covariances are not obtained), I will consider the difference in χ^2 between the intercept invariant model (if constraints are found to be tenable) and each of these models.

The decision to utilize the difference in χ^2 in these cases rests on the fact that the χ^2 difference test is a true and precise statistical test. A precise statistical test is desired in making decisions of whether or not there are significant differences in the reliable latent-construct parameters (i.e., latent variances, means, covariances, and correlations) between groups. Although testing for measurement invariance is important as a prerequisite for testing for statistical differences in the reliable latent parameters between two groups, tests of specific theoretical hypotheses needs to be based on statistical inferences (rather than practical fit). Because testing for differences or similarities in variances, means, covariances, and correlations is where research interest lies, using a more precise statistical test leads to more accurate and appropriate conclusions than using practical measures such as the CFI-difference test.

The difference in χ^2 between any two models is a test of the equality constraint placed on one model (compared to the previous model), with degrees of freedom equal to the difference in their degrees of freedom. If the test is non-significant, then the null hypothesis is not rejected indicating that there is no

difference between the 5- and 7-point scales in terms of the information gathered. If the test is significant then the null hypothesis is rejected. This result would indicate that there is a difference between the two scales (Little, 1997).

Because of the power associated with the large sample size, I will use a *p*-value of .001 as my criterion for the null hypothesis decision for omnibus tests conducted in the CFA framework. A *p*-value of .01 and .05 will be used for all univariate null hypothesis decisions that are *ad hoc* and *a priori*, respectively.

Criterion validity. In terms of criterion validity, the gold standard would be some form of alternative assessment of the objective and unbiased quality of the dimensions measured by both the long and short forms. However, for this project, such a standard does not exist. Instead, I will take an atypical approach to evaluate the criterion validity of the short-form instrument. Specifically, I will examine the short-form responses to see if factors that are known to affect student ratings also affect the short-form ratings in a similar manner. Because the current short-form instrument is not designed to remove student bias or other similar influences, it should remain sensitive to these influences even though students are responding to short-form composite items of the key dimensions of teaching quality.

For the current study, instructor reputation was measured in terms of how important it was as a reason for taking a particular course (i.e., How important were the following reasons for taking this course?). Students responded to the following "Course instructor has a good reputation" on a 4-point Likert type scale ranging from "not a reason" to "very important" (of a reason).

Class attendance was measured by asking students the following question, "Over the course of the semester, how many class meetings did you miss?" This question was asked in an open-ended fashion where students could respond from zero to 99. This information was converted to number of class hours missed (to the nearest hour) and then subtracted from the total possible class hours. Thus, class attendance was coded such that higher values reflect greater attendance and differences in how often the class met were controlled for. That is, missing two class periods in a class that meets one hour per day (to the nearest hour) on Monday, Wednesday, and Friday is only two hours whereas missing two class periods for a one-day-per-week seminar class reflects six hours of class time missed.

Expected grade was measured with the following question, "What grade do you expect in the class?" The options included twelve categories ranging from A to F (including pluses and minuses).

Construct validity. To examine construct validity of the short form, I will look at the correlations for the individual constructs in terms of the short-form composite item, the parcel of interest, and the individual items that make up both the short-form composite item and the long-form parcel of interest. If the short-form composite item for each construct is behaving reasonably well compared to the long-form parcel of interest, the correlations of the second and third parcels with the short-form composite item should be similar to the correlations of the second and third parcels with the parcel of interest from the long form. In addition, I can utilize the counterbalancing procedure to examine if the short form is susceptible to order effects

or not. This change in the administration context provides an opportunity to examine how robust the short form items are to change in the context of administration (i.e., presented first or last). Similarly, differences across classroom types such as small versus large classes, undergraduate versus graduate classes, Psychology versus non-Psychology classes, and Spring (2007) versus Summer (2007) classes can be examined to see if the short-form items behave similarly.

Results

Content validity

Means and standard deviations for the six constructs of both the short and long forms are presented in Table 3. For all analyses, short and long form responses were combined regardless of the order in which they were collected as only 3 out of 26 mean comparisons were significant at the $p = .05$ level. For the short form on the 5-point scale, only the "help" construct produced a significantly different mean difference, $t(1295) = -2.105$, $p < .05$ (a negative t statistic implying a larger mean when the short form was presented first). For the short form on the 7-point scale, there were no significant mean differences. For the long form on the 5-point scale, only the "content" construct produced a significant mean difference, $t(1295) = 2.240$, $p < .05$. Finally, for the long form on the 7-point scale, only the "goals" construct produced a significant mean difference, $t(1295) = -3.407$, $p < .01$.

Correlations between the short-form items and the long-form parcels of interest (i.e., the parcels that contain the individual items that make up the short-form construct) are presented in Table 4. Correlations between the short-form items and

their corresponding parcels of interest (i.e., correlations on the diagonal of the matrix) were found to be higher than correlations for different constructs (i.e., off-diagonal correlations) in the majority of the cases. There were certain cases in which the off-diagonal correlations were found to be higher. Overall, the correlations were in the 0.40 to 0.50 range, which is somewhat smaller than I expected, but still high enough to consider their relations in the context of the CFA model.

When comparing small and large classes (where a class of 50 or more students is considered large), significant differences in correlations were found for the short-form composite item of the "help" construct with the long-form parcel of interest of the "help" ($r_1 - r_2 = 0.12$, $p = .003$) construct, where a larger correlation was found for the smaller class size. In terms of graduate versus undergraduate courses, significant differences in correlations were found for the short-form item with the long-form parcel of interest for the "help" construct ($r_1 - r_2 = 0.13$, $p = .005$), where a larger correlation was found for graduate courses.

In terms of Spring versus Summer courses, significant differences in correlations were found for the short-form item of the "help" and "goals" constructs with the long-form parcel of interest of the "learn" ($r_1 - r_2 = 0.13$, p=.006) construct. Significant differences in correlations were also found for the "expectations" construct of the long and short forms ($r_1 - r_2 = 0.13$, $p = .004$). For all significant differences in correlations, larger correlations were found for the Summer semester.

In terms of Psychology versus non-Psychology courses, significant differences in correlations were found for the long form parcel of interest of the

"teach" construct with the short-form composite items of the "help" ($r_1 - r_2 = 0.13$, $p = .007$) and "content" ($r_1 - r_2 = 0.10$, $p = .007$) constructs. Significant differences in correlations were also found for the short-form item of the "learn" and "help" constructs with the long-form parcel of interest of the "goals" (r1 – r2 = 0.16, $p = .000$) construct. Finally, the long-form parcel of interest of the "expectations" construct produced significant differences in correlations with all short-form constructs ("teach": $r_1 - r_2 = 0.13$, $p = .005$; "learn": $r_1 - r_2 = 0.16$, $p = .000$; "help": $r_1 - r_2 = 0.13$, $p = .007$; "goals": $r_1 - r_2 = 0.14$, $p = .001$; "content": $r_1 - r_2 = 0.12$, $p = .007$; "expectations": $r_1 - r_2 = 0.141$, $p = .000$) except for the "amount learned" item. For all significant differences in correlations, larger correlations were found for non-Psychology classes.

For any two correlations being compared, Fisher's z' transformation of *r* was used (Cohen, Cohen, West, & Aiken, 2003), given by the following formula:

$$z' = 0.5[ln(1 + r) - ln(1 - r)] \quad (2)$$

Subsequently, the normal curve deviate was computed to test the null hypothesis that the difference between the population correlations is zero. The formula for this normal curve deviate is given as follows:

$$z = \frac{z'_v - z'_w}{\sqrt{1/(n_v - 3) + 1/(n_w - 3)}} \quad (3)$$

All computations were done using a software program written by Preacher (2002).

Results of the two-group CFA comparing the 5- and 7-point scales are presented in Table 5. As hypothesized, there were no significant differences found in

the factorial structure of the two scales. In other words, both loading and intercept invariance were achieved (see Table 5). Also as hypothesized, differences were found in the variances, covariances, and means between the two scales. The correlations among the constructs of the two groups were not found to be different from each other. This latter finding supports my hypothesis that differences in scales used (i.e., 5- versus 7-point) should not change the relationships among constructs.

Results of the two-group CFA comparing the 5- versus rescaled 5-point scales are presented in Table 6. As hypothesized, there were no differences in the factorial structure, variances, and covariances between the two scales. However, contrary to expectations, the means were found to be different between the 5- and rescaled 5-point scales. A possible explanation for this finding is the fact that students are more likely to endorse "5" on a 5-point scale if they agree that their instructor taught well. However, students are not as likely to endorse "7" when given a 7-point scale, even if they agree that the instructor did an excellent job. The increase in response options and the addition of the qualifier 'very much' in the 7-point scale likely led to a reduction in the willingness to endorse the highest possible response choice and thereby lowered the mean when rescaled to a 5-point scale.

The single-group CFA model containing all 6 dimensions of teaching quality (see Figure 1) was found to fit the data very well ($\chi^2_{(276,\ n=1297)} = 1529.114$, $p < .001$; RMSEA = $0.061_{\ (0.058,\ 0.063)}$; CFI = 0.991; NNFI = 0.987).

A seventh construct was added to accommodate the mean differences that were found between the 5- and rescaled 5-point scales. Every single indicator was

loaded onto this construct. This construct has a single dummy-coded variable (0 and 1) that makes the distinction between scores that are on the original 5-point metric versus scores that are on the rescaled 5-point metric. This distinction allows for interpretation of other model parameters such as correlations among the latent constructs of the six teaching evaluation dimensions, after controlling for differences in the mean structure between the two scales. Both the standardized and unstandardized loadings are presented in Table 7.

Two additional constructs were added to accommodate for the differences in the method of collection of the data (i.e., data collected on the short form versus the long form). All long-form parcels were loaded onto a long-form method construct. The loadings of the parcels onto this construct are presented in Table 8. Factor loadings of the short-form composite items onto a short-form method construct are presented in Table 9.

Loadings of both long-form parcels and short-form composite items on their respective teaching effectiveness dimensions are presented in Table 10. For each of the six dimensions, the loading of the short-form composite item appears to be smaller than the corresponding long-form parcels (after controlling for method variance between the long and short forms). However, all short-form composite items' loadings were found to be significant at the $p < .001$ level (see Table 10). Residual variances, their standard errors, and R^2 values for each indicator (of each construct) are also presented in Table 10. The residual variances and R^2 values are similar for the long-form parcels and short-form composite items. Residual variances

of the short-form composite items with their corresponding long-form parcel of interest are presented in Table 11. The correlations for the "teach" ($z = 4.74$, $p < .01$) and "help" ($z = 5.91$, $p < .01$) constructs were found to be significant, implying that unexplained (i.e., error) variability was left over (after controlling for method variability and the variability explained by their teaching effectiveness dimensions). Table 12 provides the inter-correlations among the latent constructs. All correlations were found to be significant at the $p < .01$ level and are in the 0.7 to 0.9 range. This implies that if an instructor is a good teacher, he or she will be rated highly on all dimensions of effective teaching.

In comparing the loadings for the short-form composite item and the long-form parcel of interest for each dimension (see Table 13), all comparisons were found to be significantly different from each other. These results are consistent with the results presented in Table 10 which show the loadings (beta weights) of the short-form composite items as smaller than the loadings of any parcel of the long form.

Criterion validity

Table 14 presents the correlations of the factors that have been found to influence student ratings and the short-form composite items. For instructor reputation, all correlations were significant at the $p < .01$ level which implies that the more important that an instructor's reputation was for taking the course, the more positive the ratings. The correlations range from 0.19 to 0.28 which is similar to the results found by Griffin (2001). Specifically, the positive reputation indicator

correlated 0.22 with the overall instructor rating measure and 0.20 with the overall instructor rating measure.

For class attendance, in terms of hours missed throughout the semester, the majority of correlations were not found to be statistically significant with the exception of the "learn" construct and the "amount learned" question which were found to be significant at the $p < .05$ level. This result implies that students did not rate their instructors any less favorably when they missed more hours of class. This result might lead us to conclude that perhaps students' perception of the importance of attendance and actual attendance are two different issues.

Finally, in terms of expected grade, all correlations were significant at the $p = .05$ level or lower, implying that students who expected higher grades gave their instructor more positive ratings. This was particularly the case for the "teach", "learn", and "amount learned" constructs (see Table 14).

Construct validity

The correlations for the individual constructs in terms of the short form composite item, the parcel of interest, and the individual items that make up the short form item and the long-form parcel of interest are presented in Tables 15 – 20. For all constructs, the correlation of the parcel of interest with the short-form item is higher than the correlations of the short-form item with the individual items that make up the short-form item and the long-form parcel of interest. Taking the teaching construct as an example, the correlation between the parcel of interest and the short-form composite item is 0.63 and the correlation of the short-form composite item and the

individual items that make up the short-form composite item and the parcel of interest are 0.63 for “clear”, 0.56 for “understandable”, and 0.56 for “engaging”.

For all constructs, the individual items that compose the short form correlated more highly with the long-form parcel of interest than with the short-form item. Taking the “teaching” construct as an example once again, as mentioned, the correlations of the individual items with the short form are 0.63, 0.56, and 0.56 for “clear,”, “understandable”, and “engaging” respectively. The correlations of the individual items and the long-form parcel of interest are 0.88, 0.88, and 0.81 for “clear,”, “understandable”, and “engaging” respectively. A potential explanation for the difference in magnitude between the correlations is the fact that the parcel of interest is a mathematical aggregation of a few items and would therefore contain all the method and item-specific variances resulting in inflated correlations. In any case, the short-form composite items are behaving reasonably well as reflected by medium-sized correlations (with the individual items that comprise them) that are in fact significant at the $p < .001$ level.

For each construct, the correlations of the individual items with the short form item were similar for each of the two or three items that were included in the composite short-form item. This similarity in the correlation is evidence that when students are rating a given short form item, they are taking each of the adjectives into consideration and not just focusing on one of the adjectives. Continuing with the “teaching” construct as an example, students are taking “clear” (correlated 0.63 with

the short-form item), “understandable” (correlated 0.56 with the short-form item), and “engaging (correlated 0.56 with the short-form item) into consideration.

Additional construct validity analyses

When comparing small and large classes, where a class of 50 or more students was considered large, significant mean differences were found for the “learn” ($t(1221) = 3.26$, $p = .001$) and “help” ($t(1221) = 5.23$, $p = .000$) short-form composite items. The positive t-statistic values imply that students rated their instructors higher on these two dimensions in smaller classes. In terms of graduate versus undergraduate courses, no significant mean differences were found at the $p = .01$ level. In terms of Spring versus Summer courses, significant mean differences were found for items of all dimensions. The t-statistic values for all dimensions are as follows: $t(1221) = -3.65$, $p = .000$ for the “teach” item, $t(1221) = -4.17$, $p = .000$ for the “learn” item, $t(1221) = -4.36$, $p = .000$ for the “help” item, $t(1221) = -2.80$, $p = .005$ for the “goals” item, $t(1221) = -3.42$, $p = .001$ for the “content” item, and $t(1221) = -3.43$, $p = .001$ for the “expectations” item. The results found for Spring versus Summer courses imply students rated their instructors higher on items of all dimensions in the Summer. Finally, in terms of Psychology versus non-Psychology courses, a significant mean difference was found for the “content” ($t(1221) = 3.067$, $p = .002$) short-form composite item. For this particular construct, a higher mean was found for Psychology courses compared to non-Psychology courses.

Discussion

The analyses presented with regard to content, criterion, and construct validity lend some support to the overall validity of the short-form composite items. The results presented for both the short and long forms suggest that the long form is a reliable measure that performed well in the CFA analyses which supports its validity and the short form is a reasonable facsimile of the long form.

Based on these findings and the need to move forward quickly, the task force on the assessment of teaching and learning has recommended a short form comparable to the short form tested as part of my dissertation. This recommended form includes an additional item reflecting classroom climate. In addition, there are some minor differences between the nine key question of this form (Table 21) and the nine key questions on the original form (Appendix C). This modified form will be used beginning in the Fall of 2008.

Paying closer attention to the correlations of the short-form composite items and the long form parcels of interest, there does not seem to be a very clear pattern to justify several different constructs of teaching performance. That is, all dimensions of teaching and learning are highly positively correlated, suggesting that good teachers are rated positively on all dimensions. This high correlation was evident in both the long and short forms and therefore appears to be a general issue across both forms. In this regard, the issue may be specific to the sample of instructors used in the current study. That is, the instructors who agreed to be a part of this study may be more uniformly better or worse along these dimensions than the general population of

instructors. However, more detailed analyses and a larger sample of instructors, courses, and areas of study may begin to reveal unique patterns in the different dimensions.

Despite the fact that there does not seem to be a very clear pattern to justify several different constructs of teaching performance, having several dimensions that tap into different aspects of teaching and perceptions of student learning in a relatively short format seems justifiable. This approach allows instructors to identify possible deficits in specific areas of teaching quality – a benefit that does not exist by using a single overall rating of teaching quality.

Additional construct validity analyses

The additional analyses performed comparing students' ratings on the short-form composite items for small versus large classes revealed significant mean differences for the "learn" and "help" constructs. This implies that students felt that they learned more and that help was more easily attainable in smaller classes, but that their instructor's teaching ability was not influenced by class size.

In terms of Spring versus Summer classes, significant mean differences were found for all dimensions of teaching effectiveness. Potential reasons for higher ratings in the Summer could be due to (a) an instructor being more focused on his or her class as there are not the usual school year activities occurring simultaneously, (b) instructor who teach in the summer enjoy teaching more than other instructors, (c) students being more focused on a particular course (as opposed to having to balance

their time among 5 courses, for example), and (d) students being able to retain more information as exams occur more frequently during the Summer semester.

Limitations

One of the key limitations of the current study is the lack of a representative sample of instructors, courses, and area of study. The sample was a convenience sample and may, therefore, be selective in nature. This potential selectivity is reflected in the majority of classes being Psychology classes (23 out of 51 classes). Furthermore, some instructors had up to 3 sections. For example, one instructor had 2 sections in the Spring semester and one section in the Summer semester. An instructor being represented more than once in the sample further adds to the selectively of the sample used in the current study.

Another limitation of the current study is the lack of negatively worded items (e.g., an adjective such as 'confusing'); it does not make logical sense to have both a positively and negatively worded items incorporated (in a single sentence) for the short-form composite item case. Generally speaking, including negatively worded items is beneficial because having negatively worded items can help to reduce response-set biases that can occur with all items worded in the same (positive) direction.

Finally, the most important limitation of the current study is that a gold standard criterion to gauge the quality of the short-form items does not exist. Short-form responses are simply being compared to those of the long form. However, the

long form items have not been extensively validated and their data are derived from the same, potentially selective, sample.

The degree to which these limitations will bias the final results is unknown. On the one hand, the long and short forms show some evidence of being valid in spite of these limitations. In fact, the limitations may be contributing to the lack of differentiation among the teaching constructs assessed by both forms. Clearly, future work is needed to further examine the utility and validity of a short form of teaching and learning.

Future Directions

As I move forward in my career, I plan to continue research in the general area of teaching and student learning. In the following section, I outline some directions I expect to follow both in terms of (a) teaching evaluations and (b) establishing a gold standard to measure teaching effectiveness and student learning.

Establishing a happy medium. Ideally, I would like to create a teaching evaluation tool that is a happy medium between the long (Appendix B) and short (Appendix C) forms. This particular teaching evaluation tool will look very similar to the long form. However, each construct will likely have 4 items (versus the 10 – 12 items in the current long form) – 2 positively worded and 2 negatively worded items. The specific items that make up this "happy medium" teaching evaluation protocol will likely be obtained from the current long form, but based on data from a larger representative sample of instructors. An example of this presentation format for this medium-form scale is given in Table 22.

One of the primary benefits of this compromise approach is the fact that students have the option of rating each adjective individually. Therefore, a student has the ability to rate his or her instructor's teaching on the dimensions of 'clear', 'organized' as well as 'confusing' and 'detached', for example. These individual judgments are not possible in the composite item case. Another advantage of this medium approach (as outlined above) is the possibility of having both positively- and negatively-worded items for each construct. This use of positive and negative items is (as mentioned) important because (a) it serves as a check that students are not simply selecting the same response for every single item (and are not paying attention to the actual questions) and (b) it allows us to treat our existing constructs as latent constructs, with each construct having positive and negative facets and these facets having 2 or 3 indicators each.

This medium-length protocol might take a longer amount of time to respond to compared to responding to the short-form composite items. However, it should not take too much longer because all judgments for each construct have the same beginning structure (e.g., "The instructor's teaching was…" for the "teaching" construct). Even if this 'happy medium' instrument takes a little bit more time to fill out compared to the current short form protocol, the potential benefits (e.g., the ability to calculate internal consistency estimates, reduce response sets,) may prove to be an improvement on the composite single-item approach. Of course, the validity of such an approach would have to be tested to determine that it does provide an improvement over the single-item approach examined as part of my dissertation.

Using item response theory to identify more or less discriminating items. Another idea I have in terms of teaching evaluations involves using item response theory (IRT) to identify more or less discriminating items for each construct. IRT is a relatively new measurement system that is an alternative to classical test theory (CTT). IRT is commonly used in large testing companies in the United States and Europe for design of tests and construction of test item banks, among other purposes (Hambleton, Swaminathan, & Rogers, 1991).

One of the key limitations of CTT, in the context of teaching evaluations, is that student characteristics and the characteristics of the teaching evaluation tool cannot be separated (i.e., one can be interpreted only within the context of the other). The student characteristic of interest is a "trait" that can be conceptualized as an attitude toward the instructor and the course. In the CTT framework, this "trait" is defined only in terms of one specific teaching evaluation tool (i.e., the protocol used in the current study). Essentially, characteristics of the protocol and its items influence student characteristics and student characteristics in turn influence the evaluation tool (Hambleton et al., 1991). For this reason, it becomes difficult to compare students who respond to different evaluation tools (i.e., the evaluation instrument already used versus data that will be collected on a slightly different short-form in Fall 2008) and to compare items whose information is obtained from different groups of students.

IRT provides a framework where item characteristics are not group-dependent and responses obtained by students are not dependent on a specific teaching

evaluation tool. The more popular IRT models are dichotomous in nature and have the capability of adjusting for specific properties of items such as their difficulty (i.e., how much of the trait is required for endorsement of an item), discrimination (i.e., how quickly the probability level changes for a unit increase in the trait), and tendency toward guessing. The guessing parameter was proposed by Birnbaum specifically for the purposes of the three-parameter logistic model to account for the nonzero performance of low-ability examinees on multiple-choice items (van der Linden & Hambleton, 1996). Since the "trait" of interest, in the context of teaching evaluations, is the student's attitude toward the instructor and the course, this guessing parameter does not apply and would therefore not be reflected in the model outlined below.

The model I will be considering is the graded response model, which consists of a family of mathematical models that deals with ordered polytomous categories (Samejima, 1996). Examples of ordered polytomous categories include letter grades used in the evaluation of student performance and student responses on a 5-point Likert-type scale used to evaluate teaching performance.

The graded response model will allow me to identify more discriminating (i.e., "difficult") or less discriminating (i.e., "easy") items of a teaching evaluation tool depending on the question of interest. For example, it is possible to identify a subset of "easy" items for the instructor's "teaching" construct if we are trying to discover whether an instructor is adequate (i.e., whether or not an instructor meets a minimum threshold of teaching quality). In addition, it is possible to identify an

entirely different subset of "difficult" items if the intent is to make a distinction between a good and an excellent instructor for this same construct.

In general, the category response function of the graded response model is given by the following:

$$P_{ui}(\theta) \equiv \text{Prob}[U_i = u_i \mid \theta] \tag{4}$$

where $P_{ui}(\theta)$ is the probability with which a student with attitude θ gives a response u_i to item i. U_i is a random variable used to denote the graded item response to item i. The category response function of the logistic model is given by the following:

$$P_{ui}(\theta) = \frac{\exp[-Da_i(\theta - b_{ui+1})] - \exp[-Da_i(\theta - b_{ui})]}{\{1 + \exp[-Da_i(\theta - b_{ui})]\}\{1 + \exp[-Da_i(\theta - b_{ui+1})]\}} \tag{5}$$

where D = 1.7 is a scaling factor that is introduced to make the logistic model more similar to the normal ogive (i.e., cumulative normal distribution) model, a_i is the item discrimination parameter, and b_{ui} is the item difficulty parameter. PARSCALE (Muraki & Bock, 1993) can be used to estimate the item parameters for item discrimination (i.e., a_i) and item difficulty (i.e., b_{ui}).

Establishing a gold standard. The recommendations put forth by the task force on the assessment of teaching and learning consider more than one form of assessing teaching effectiveness and student learning. However, all task force recommendations involve either student or peer feedback, both of which might be biased in nature. As discussed, students tend to be biased by the grade that they expect and by how important instructor reputation is as a reason for taking a particular class. Similarly, peers might be biased for their own benefit. That is, they

might not rate an instructor as positively as they might deserve to be rated so that they themselves will look better in terms of their evaluations. Perhaps an unbiased, objective method of evaluating teaching effectiveness and student learning needs to be proposed.

One possible way of conducting an unbiased evaluation of teaching effectiveness and student learning would involve training assistants (perhaps graduate students who are interested in the scholarship of teaching and learning) to conduct observations in actual instructors' classrooms. A coding schema could be developed to tap into the dimensions that make up effective teaching such as organization and clarity. A similar coding rubric can be developed for assessing student learning. For example, these assistants can code how engaged students are during a lecture or discussion session as well as review the work conducted by students.

An additional coding rubric tapping into important teaching and learning aspects that occur outside of the classroom can be developed. For example, these trained assistants can interview both students and instructors regarding issues such as obtaining assistance with the course materials outside of the regular classroom. Discrepancies between instructor and student interviews that arise can be resolved by looking at more objective measures (e.g., e-mails, visits to office hours) of out-of-classroom interactions.

All coded observations and interviews can then be correlated with both student- and peer-report measures of teaching effectiveness and student learning. This objective observation and interviewing process could potentially be carried out every

3 years or so for each full-time instructor. The initial implementation of this "gold standard" (in terms of developing the necessary rubrics and training graduate student assistants) might prove to be rather time-consuming. However, having yet another method of evaluating both teaching and student learning that does not appear to be biased in nature is something that I feel is worth pursuing.

Conclusions

Based on the results found in the current study, there is sufficient evidence to justify the utilization of short-form composite items. Specifically, this justification is based on the evidence found for the three types of validity. In term of content validity, focusing our attention on the confirmatory factor analysis results, the factor loadings for all short-form items were found to be significant, despite their smaller magnitude in comparison to the long-form parcels for each construct of interest.

In terms of criterion-based validity, for the most part, the factors (i.e., instructor reputation, expected grade, and class attendance) that were found to influence students' responses to the short-form composite items were also found in previous studies reviewed. Finally, in terms of construct validity, the correlations of the individual items that comprise the short-form composite item were significant at the $p < .001$ level and were fairly close to one another in terms of magnitude. As mentioned, this is evidence that students are not just taking one adjective into consideration and ignoring the other two, but are in fact taking all adjectives that comprise a composite item into consideration.

The evidence found for the different types of validities provide different perspectives on how and why short-form composite item are valid. According to Messick, it is not an issue of whether an item is valid or not, but rather an issue of the degree of validity (Messick, 1989). As one might expect, the long form with multiple items per construct appears to behave better than the short-form items. However, if one recalls the primary question of how much damage would result from using a short form, then the answer appears to be "not much at all".

Final thoughts

The current project seeks to demonstrate the validity of a short form in the context of teaching evaluations. As mentioned, there is plenty more that can be done with regard to teaching evaluations. Specifically, I would like to pursue the "happy medium" approach if feasible. In trying to make teaching evaluations more concise and focused on what students are able to judge best, one should not lose sight of the fact that student feedback is only one of the methods of tapping into teaching effectiveness. For this reason, it seems important to pursue the creation and development of a gold standard for measuring teaching effectiveness. Until further work can be done, I would conclude that the current short form is an adequate measure of students' perceptions of instructor effectiveness. Future work and follow-up studies will need to be conducted to determine if the current approach can be improved to an appreciable degree.

Endnotes

[1] The idea that the sum of indicators is more reliable than individual items can be understood by considering the classical measurement theorem:

$$X_i = T_i + S_i + e_i$$

where X_i represents the score of an individual for a particular item, T_i represents the reliable true score variance, S_i represents the item specific variance, and e_i represents random error (i.e., unreliability of the item). From this classical test theory (CTT) perspective, the response by an individual to a given observed item is composed of three sources of variance: a 'true' core aspect (i.e., the part of an item that assesses the construct we desire to measure), a 'specific' component (i.e., a reliable component that is specific to the item, but is unrelated to the construct", and a random error component depicting the random noise that exists in the measurement process (Widaman et al., 2008). As more items are added to tap into a particular construct, there is a larger contribution of the 'true' core aspect of variability (since every item contributes to this component), relative to S_i and e_i, which are uncorrelated sources of variance across items measuring a particular construct. As a result, there is an increase in the ratio of true score variance to total observed variance, which is the definition of reliability.

[2] The "help" construct was considered as two separate constructs with the following stems: "When I asked for help, this instructor was:" and "Although I did not seek for help, this instructor indicated that s/he would be:" If students simultaneously

responded to both sections, the assumption was made that they did in fact seek help from their instructor.

[3] An “amount learned” question was included for the short form but is not represented in the long form. This “amount learned” question was as follows: “Compared to other similar courses, I would rate how much I learned as:” and the responses, on a 5-point scale were as follows: “much below expected”, “below expected”, “expected/average”, “above expected”, and “much above expected”.

[4] The question pertaining to the amount a student learned compared to other similar courses is included for the short form.

References

Amin, M. E. (2002). Six factors of course and teaching evaluation in a bilingual university in Central Africa. *Assessment and Evaluation in Higher Education, 27*, 281-291.

Bausell, R. B., & Magoon, J. (1972). Expected grade in a course, grade point average, and student ratings of the course and the instructor. *Educational and Psychological Measurement, 32*, 1013-1023.

Blackhart, G. C., Peruche, M., DeWall, C. N., & Joiner, T. E., Jr. (2006). Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*, 37-39.

Blum, M. L. (1936). An investigation of the relation existing between students' grades and their rating of the instructors' ability to teach. *Journal of Educational Psychology, 27*, 217-221.

Brown, D. L. (1976). Faculty-ratings and student grades: A university-wide multiple regression analysis. *Journal of Educational Psychology, 68*, 573-578.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

Burns, S. M., & Ludlow, L. H. (2005). Understanding student evaluations of teaching quality: The contributions of class attendance. *Journal of Personnel Evaluation in Education, 18*, 127-138.

Byrne, C. J. (1992). Validity studies of teacher rating instruments: design and interpretation. *Research in Education, 48*, 42-54.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd edition).* Mahwah, NJ: Erlbaum.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Garverick, C. M., & Carter, H. D. (1962). Instructor ratings and expected grades. *California Journal of Educational Research, 13*, 218-221.

Ginexi, E. M. (2003). General psychology course evaluations: Differential survey response by expected grade. *Teaching of Psychology, 30*, 248-251.

Griffin, B. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*, 534-552.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundametals of item response theory.* Newbury Park, CA: Sage.

Hill, T., & Lewicki, P. (2005). *Statistics: Methods and Applications: A comprehensive reference for science, industry, and data mining.* StatSoft, Inc.

Holmes, D. S. (1971). The relationship between expected grades and students' evaluations of their instructors. *Educational and Psychological Measurement, 31*, 951-957.

Husbands, C. T., & Fosh, P. (1993). Students' evaluation of teaching in higher education: Experiences from four different European countries and some

implications of the practice. *Assessment and Evaluation in Higher Education, 18*, 95-115.

Kishton, J. M., Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757-765.

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Educational Review, 18*, 59-63.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151 – 173.

Marlin, J. E., Jr. & Gaynor, P. (1989). Do anticipated grades affect student evaluations? *College Student Journal, 23*, 184-192.

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33*, 176-179.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in test of measurement invariance. *Journal of Applied Psychology, 93*, 568-592.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd edition; pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Preacher, K. J. (2002, May). *Calculation for the test of the difference between two independent correlation coefficients* [Computer software]. Available from http://www.quantpsy.org.

Samejima, F. (1996). Graded response model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, pp. 85 – 100, New York, NY: Springer.

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: individual versus class effects. *Journal of Educational Psychology, 71*, 293-302.

van der Linden, W. J. & Hambleton, R. K. (1996). *Handbook of Modern Item Response Theory.* New York, NY: Springer.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2008). On creating and using short forms of scales in archival research. In K. Trzesniewski, Donnellen, & Lucas (Eds.), *Obtaining and analyzing archival data: Methods and illustrations.* Washington, DC: American Psychological Association (under review).

Youngblut, J. M., & Casper, G. R. (1993). Focus on psychometrics: Single-item indicators in nursing research. *Research in Nursing & Health*, *16*, 459-465.

Table 1

Breakdown of classes used in the initial validation of the new teaching evaluation tool

Department	Class number	Line number	Semester & Year	n	G vs. UG
PSYC	104	58859	Spring 2007	40	UG
PSYC	104	58860	Spring 2007	35	UG
PSYC	104	88867	Summer 2007	27	UG
PSYC	104	67069	Spring 2007	36	UG
PSYC	104	67070	Spring 2007	31	UG
PSYC	300	58866	Spring 2007	33	UG
PSYC	300	58868	Spring 2007	25	UG
PSYC	300	62723	Spring 2007	29	UG
PSYC	300	63686	Spring 2007	15	UG
PSYC	300	93251	Summer 2007	14	UG
PSYC	333	88883	Summer 2007	17	UG
PSYC	333	58872	Spring 2007	72	UG
PSYC	333	91271	Summer 2007	13	UG

Department	Class number	Line number	Semester & Year	n	G vs. UG
PSYC	350	58873	Spring 2007	208	UG
PSYC	350	88887	Summer 2007	10	UG
PSYC	360	58874	Spring 2007	19	UG
PSYC	360	58875	Spring 2007	15	UG
PSYC	370	88893	Summer 2007	9	UG
PSYC	430	68809	Spring 2007	50	UG
PSYC	465	91547	Summer 2007	21	UG
PSYC	535	58925	Spring 2007	31	UG
PSYC	850	68928	Spring 2007	6	G
PSYC	896	12345	Spring 2007	31	G
COMS	130	91155	Summer 2007	14	UG
COMS	130	93653	Summer 2007	17	UG
COMS	130	82823	Summer 2007	13	UG
COMS	235	91157	Summer 2007	24	UG
COMS	246	94863	Summer 2007	9	UG
COMS	330	82843	Summer 2007	13	UG

Department	Class number	Line number	Semester & Year	n	G vs. UG
COMS	330	82841	Summer 2007	8	UG
COMS	330	82839	Summer 2007	11	UG
COMS	331	82845	Summer 2007	9	UG
COMS	342	82851	Summer 2007	24	UG
COMS	955	69298	Spring 2007	11	G
PRE	598	58598	Spring 2007	13	UG
PRE	725	58605	Spring 2007	15	G
PRE	811	58608	Spring 2007	32	G
PRE	835	68897	Spring 2007	9	G
PRE	903	61906	Spring 2007	24	G
PRE	998	70104	Spring 2007	5	G
PRE	998	70105	Spring 2007	4	G
C&T	840	70699	Spring 2007	16	G
C&T	896	69945	Spring 2007	18	G
HSES	244	62670	Spring 2007	35	UG
HSES	350	66599	Spring 2007	35	UG

Department	Class number	Line number	Semester & Year	n	G vs. UG
HSES	529	55180	Spring 2007	15	UG
HSES	654	55201	Spring 2007	15	UG
HSES	675	55211	Spring 2007	53	UG
ELPS	755	60239	Spring 2007	9	G
ELPS	955	60382	Spring 2007	11	G
SPED	785	68399	Spring 2007	20	G

Note: PSYC = Psychology; COMS = Communication Studies; PRE = Psychology and Research in Education; C&T = Curriculum and Training; HSES = Health, Sport, and Exercise Sciences; ELPS = Educational Leadership and Policy Studies; SPED = Special Education; G vs. UG = Graduate versus Undergraduate course

Table 2

Cronbach's α values for long form constructs on the 5- and 7-point scales

	5-point	7-point
Constructs	Cronbach's α	Cronbach's α
Teaching	0.924	0.928
Learning	0.873	0.934
Help	0.938	0.943
Goals	0.883	0.852
Content	0.928	0.925
Expectations	0.823	0.828

Table 3

Means and standard deviations for the short- and long-form constructs

		Short form		Long form	
Construct	n	Mean	SD	Mean	SD
Teaching	1297	4.423	0.801	4.387	0.724
Learning	1297	4.412	0.783	4.318	0.762
Help	1297	4.383	0.805	4.417	0.744
Goals	1297	4.441	0.766	4.464	0.760
Content	1297	4.422	0.805	4.291	0.802
Expectations	1297	4.374	0.812	4.227	0.800
Amount learned	1297	3.781	0.782	---	---

Table 4

Correlations between the long and short form constructs

		Long form					
		Teaching	Learning	Help	Goals	Content	Expectations
Short form	Teaching	0.631	0.492	0.456	0.521	0.546	0.436
	Learning	0.557	0.598	0.557	0.542	0.546	0.454
	Help	0.412	0.494	0.581	0.442	0.475	0.382
	Goals	0.555	0.537	0.500	0.596	0.608	0.461
	Content	0.612	0.527	0.512	0.571	0.604	0.441
	Expectations	0.539	0.523	0.478	0.548	0.548	0.519
	Amount learned	0.479	0.403	0.348	0.385	0.428	0.347

Note: All correlations are significant at the $p<.0001$ level

Table 5

Two-group CFA results (5- versus 7-point scales)

Model	$\chi 2$	*df*	*p*	$\Delta\chi 2$	Δ *df*	*p*	RMSEA	90% C.I.	NNFI	CFI	ΔCFI	Constraint Tenable
Configural	2071.662	576	<.001	---	---		.083	.079-.087	0.977	0.981	---	---
Loading invariance	2172.411	597	<.001	---	---		.085	.082-.089	0.977	0.980	0.001	Yes
Intercept invariance	2328.677	618	<.001	---	---		.086	.082-.089	0.979	0.979	0.001	Yes
Homogeneity of Var/Covar	2375.467	633	<.001	49.799	21	<.001	.086	.083-.090	0.976	0.978	---	No
Variances only*	2988.533	630	<.001	659.856	12	<.001	.105	.102-.108	0.967	0.971	---	No
Means*	2601.143	624	<.001	272.466	6	<.001	.089	.086-.093	0.972	0.975	---	No
Correlations only*	2353.874	632	<.001	25.197	14	.033	.085	.082-.089	0.976	0.979	---	Yes

*These models are compared to the intercept invariance model and not the model that precedes it.

Table 6

Two-group CFA results (5- versus rescaled 5-point scales)

Model	χ2	*df*	*p*	Δ χ2	Δ *df*	*p*	RMSEA	90% C.I.	NNFI	CFI	ΔCFI	Constraint Tenable
Configural	2878.905	576	<.001	---	---	---	.088	.085-.090	0.979	0.983	---	---
Loading invariance	2934.919	597	<.001	---	---	---	.087	.084-.089	0.980	0.983	0.000	Yes
Intercept invariance	3021.077	618	<.001	---	---	---	.087	.084-.090	0.980	0.982	0.001	Yes
Homogeneity of Var/Covar	3035.019	633	<.001	13.936	15	.53	.086	.083-.089	0.980	0.982	---	Yes
Means*	3069.898	624	<.001	48.821	6	<.001	.087	.084-.090	0.980	0.982	---	No

*This model is compared to the intercept invariance model and not the model that precedes it.

Table 7

Lambda loadings of the scale construct to accommodate for mean differences between 5- and rescaled 5-point scales

Indicator	Unstandardized loadings	Standard error	*p*	Standardized loadings
Teaching				
Parcel 1	-0.177	0.040	<.01	-0.124
Parcel 2	-0.068	0.041	ns	-0.046
Parcel 3	-0.084	0.042	<.05	-0.055
Parcel 4	-0.098	0.037	<.01	-0.074
Composite item	-0.160	0.040	<.01	-0.110
Learning				
Parcel 1	-0.156	0.042	<.01	-0.103
Parcel 2	-0.099	0.040	<.05	-0.069
Parcel 3	-0.067	0.042	ns	-0.045
Parcel 4	-0.091	0.040	<.05	-0.062
Composite item	-0.148	0.040	<.01	-0.103
Help				

Indicator	Unstandardized loadings	Standard error	*p*	Standardized loadings
Parcel 1	-0.083	0.041	<.05	-0.056
Parcel 2	-0.022	0.042	ns	-0.015
Parcel 3	-0.011	0.041	ns	-0.007
Composite item	-0.081	0.042	ns	-0.053
Goals				
Parcel 1	-0.153	0.042	<.01	-0.102
Parcel 2	-0.229	0.046	<.01	-0.137
Parcel 3	-0.291	0.046	<.01	-0.174
Parcel 4	-0.096	0.045	<.05	-0.059
Composite item	-0.141	0.039	<.01	-0.101
Content				
Parcel 1	-0.144	0.044	<.01	-0.091
Parcel 2	-0.078	0.042	ns	-0.052
Parcel 3	-0.012	0.043	ns	-0.008
Composite item	-0.100	0.041	<.05	-0.069

Indicator	Unstandardized loadings	Standard error	*p*	Standardized loadings
Expectations				
Parcel 1	-0.126	0.044	<.01	-0.079
Parcel 2	-0.059	0.046	ns	-0.036
Parcel 3	-0.074	0.040	<.01	-0.051
Composite item	-0.122	0.041	<.01	-0.082

Table 8

Lambda loadings of the long-form method construct

Indicator	Unstandardized loadings	Standard error	*p*	Standardized loadings
Teaching				
Parcel 1	-0.115	0.026	<.01	-0.161
Parcel 2	0.120	0.027	<.01	0.161
Parcel 3	0.105	0.027	<.01	0.138
Parcel 4	0.041	0.024	ns	0.062
Composite item	---	---	---	---
Learning				
Parcel 1	-0.074	0.028	<.05	-0.097
Parcel 2	0.130	0.026	<.01	0.180
Parcel 3	0.113	0.028	<.01	0.149
Parcel 4	0.119	0.027	<.01	0.164
Composite item	---	---	---	---
Help				

Indicator	Unstandarized loading	Standard error	*p*	Standarized loading
Parcel 1	-0.130	0.027	<.01	-0.176
Parcel 2	0.106	0.028	<.01	0.141
Parcel 3	0.070	0.028	<.01	0.094
Composite item	---	---	---	---
Goals				
Parcel 1	-0.120	0.027	<.01	-0.159
Parcel 2	0.243	0.030	<.01	0.289
Parcel 3	0.145	0.030	<.01	0.173
Parcel 4	0.250	0.030	<.01	0.308
Composite item	---	---	---	---
Content				
Parcel 1	-0.093	0.030	<.01	-0.116
Parcel 2	0.162	0.028	<.01	0.215
Parcel 3	0.182	0.029	<.01	0.233
Composite item	---	---	---	---

Indicator	Unstandarized loading	Standard error	*p*	Standarized loading
Expectations				
Parcel 1	-0.068	0.029	<.05	-0.086
Parcel 2	0.313	0.029	<.01	0.378
Parcel 3	0.058	0.027	<.05	0.080
Composite item	---	---	---	---

Table 9

Lambda loadings of the short-form method construct

Construct	Unstandardized loading	Standard Error	*p*	Standardized loading
Teaching	0.464	0.015	<.01	0.636
Learning	0.439	0.015	<.01	0.612
Help	0.407	0.017	<.01	0.534
Goals	0.487	0.014	<.01	0.696
Content	0.499	0.015	<.01	0.682
Expectations	0.477	0.016	<.01	0.638

Table 10

Lambda loadings, residuals, and R^2 values of the single-group CFA model

Indicator	Unstandardized loadings	Standard error	*p*	Residuals	Standard error	R^2
Teaching						
Parcel 1	0.586	0.017	<.001	0.151	0.008	0.707
Parcel 2	0.635	0.017	<.001	0.136	0.007	0.756
Parcel 3	0.611	0.018	<.001	0.187	0.009	0.673
Parcel 4	0.567	0.015	<.001	0.112	0.006	0.743
Composite item	0.376	0.017	<.001	0.169	0.008	0.682
Learning						
Parcel 1	0.639	0.017	<.001	0.156	0.008	0.729
Parcel 2	0.611	0.017	<.001	0.128	0.006	0.754
Parcel 3	0.627	0.018	<.001	0.164	0.008	0.713
Parcel 4	0.614	0.017	<.001	0.138	0.007	0.740
Composite item	0.405	0.017	<.001	0.154	0.008	0.702
Help						

Indicator	Unstandardized loadings	Standard error	*p*	Residuals	Standard error	R^2
Parcel 1	0.653	0.018	<.001	0.099	0.007	0.818
Parcel 2	0.673	0.017	<.001	0.099	0.006	0.825
Parcel 3	0.671	0.016	<.001	0.100	0.006	0.820
Composite item	0.366	0.019	<.001	0.278	0.012	0.520
Goals						
Parcel 1	0.644	0.018	<.001	0.132	0.009	0.767
Parcel 2	0.569	0.022	<.001	0.309	0.014	0.562
Parcel 3	0.551	0.021	<.001	0.354	0.015	0.494
Parcel 4	0.627	0.021	<.001	0.201	0.010	0.696
Composite item	0.389	0.016	<.001	0.096	0.006	0.804
Content						
Parcel 1	0.670	0.018	<.001	0.172	0.010	0.729
Parcel 2	0.661	0.018	<.001	0.106	0.006	0.814
Parcel 3	0.665	0.019	<.001	0.136	0.007	0.778
Composite item	0.408	0.017	<.001	0.118	0.007	0.780
Expectations						

Indicator	Unstandardized loadings	Standard error	*p*	Residuals	Standard error	R^2
Parcel 1	0.606	0.020	<.001	0.260	0.013	0.592
Parcel 2	0.593	0.023	<.001	0.236	0.013	0.656
Parcel 3	0.583	0.017	<.001	0.183	0.009	0.652
Composite item	0.409	0.018	<.001	0.160	0.008	0.714

Table 11

Correlated residuals of short-form composite items and their corresponding long-form parcel of interest

Teaching	Learning	Help	Goals	Content	Expectations
0.028	0.011	0.037	-0.009	-0.006	0.012
(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.008)
z=4.740, p<.01	z=1.944, ns	z=5.905, p<.01	z=-1.745, ns	z=-1.058, ns	z=1.620, ns

Table 12

Correlations among latent constructs of the single-group CFA (ψ estimates)

	Teaching	Learning	Help	Goals	Content	Expectations
Teaching	1.000					
Learning	0.836	1.000				
	(0.011)					
	z=74.356, p<.01					
Help	0.735	0.862	1.000			
	(0.015)	(0.010)				
	z=47.902, p<.01	z=88.268, p<.01				
Goals	0.761	0.814	0.788	1.000		
	(0.015)	(0.013)	(0.014)			
	z=50.426, p<.01	z=63.625, p<.01	z=57.581, p<.01			
Content	0.804	0.823	0.762	0.906	1.000	
	(0.013)	(0.012)	(0.014)	(0.009)		
	z=63.213, p<.01	z=69.793, p<.01	z=53.388	z=104.518, p<.01		
Expectations	0.744	0.810	0.738	0.884	0.869	1.000
	(0.017)	(0.014)	(0.017)	(0.011)	(0.011)	
	z=44.362, p<.01	z=57.978	z=44.034, p<.01	z=77.956, p<.01	z=76.411, p<.01	

Table 13

Lambda loading comparisons of the short-form composite items and their corresponding long form parcel of interest

Model	$\chi 2$	*df*	*p*	$\Delta \chi 2$	Δ *df*	*p*	Constraint Tenable
Single-group CFA	1529.114	276	<.001	---	---	---	---
Teaching*	1643.417	277	<.001	114.303	1	<.001	No
Learning*	1669.323	277	<.001	140.209	1	<.001	No
Help*	1737.035	277	<.001	207.921	1	<.001	No
Goals*	1682.922	277	<.001	153.808	1	<.001	No
Content*	1680.541	277	<.001	151.427	1	<.001	No
Expectations*	1598.713	277	<.001	69.599	1	<.001	No

*These models are compared to the original single-group CFA model and not the model that precedes it.

Table 14

Factors that influence student ratings

	Instructor reputation	Hours missed	Expected grade
Teaching	0.221**	-0.036	0.101**
Learning	0.200**	-0.046	0.118**
Help	0.185**	-0.023	0.052
Goals	0.190**	-0.006	0.071*
Content	0.186**	-0.028	0.077**
Expectations	0.207**	-0.026	0.086**
Amount learned	0.276**	-0.089**	0.177**

Note: * All correlations are significant at the p<.05 level
** All correlations are significant at the p<.01 level

Table 15

Correlations of the "teach" construct

	Composite	Parcel 1	Parcel 2	Parcel 3	Parcel 4	Item 1	Item 3	Item 4
Composite	1.000							
Parcel 1	0.631	1.000						
Parcel 2	0.508	0.662	1.000					
Parcel 3	0.524	0.645	0.743	1.000				
Parcel 4	0.316	0.422	0.492	0.454	1.000			
Item 1	0.627	0.879	0.620	0.579	0.398	1.000		
Item 3	0.563	0.883	0.587	0.540	0.384	0.384	1.000	
Item 4	0.561	0.814	0.598	0.645	0.402	0.633	0.624	1.000

Note: * All correlations are significant at the p<.0001 level
Item 1 = clear; Item 3=understandable; Item 4=engaging

Table 16

Correlations of the "learn" construct

	Composite	Parcel 1	Parcel 2	Parcel 3	Parcel 4	Item 5	Item 6	Item 11
Composite	1.000							
Parcel 1	0.598	1.000						
Parcel 2	0.531	0.696	1.000					
Parcel 3	0.560	0.699	0.725	1.000				
Parcel 4	0.364	0.456	0.525	0.518	1.000			
Item 5	0.555	0.905	0.654	0.629	0.429	1.000		
Item 6	0.587	0.889	0.687	0.727	0.446	0.796	1.000	
Item 11	0.489	0.818	0.534	0.558	0.363	0.658	0.610	1.000

Note: * All correlations are significant at the p<.0001 level
Item 5=encouraging; Item 6=supportive; Item 11=involved

Table 17

Correlations of the "help" construct

	Composite	Parcel 1	Parcel 2	Parcel 3	Item 1	Item 3	Item 7
Composite	1.000						
Parcel 1	0.581	1.000					
Parcel 2	0.491	0.163	1.000				
Parcel 3	0.503	0.786	0.824	1.000			
help1	0.542	0.884	0.656	0.713	1.000		
help3	0.568	0.882	0.710	0.704	0.710	1.000	
help7	0.520	0.912	0.775	0.771	0.755	0.774	1.000

Note:* All correlations are significant at the p<.0001 level
Item 1=available; Item 3=responsive; Item 7=helpful

Table 18

Correlations of the "goals" construct

	Composite	Parcel 1	Parcel 2	Parcel 3	Parcel 4	Item 1	Item 3
Composite	1.000						
Parcel 1	0.596	1.000					
Parcel 2	0.401	0.518	1.000				
Parcel 3	0.393	0.548	0.657	1.000			
Parcel 4	0.341	0.420	0.396	0.287	1.000		
Item 1	0.587	0.922	0.504	0.524	0.449	1.000	
Item 3	0.547	0.898	0.496	0.523	0.383	0.742	1.000

Note: * All correlations are significant at the p<.0001 level
Item 1=clear; Item 3=appropriate

Table 19

Correlations of the "content" construct

	Composite	Parcel 1	Parcel 2	Parcel 3	Item 6	Item 2	Item 5
Composite	1.000						
Parcel 1	0.604	1.000					
Parcel 2	0.565	0.718	1.000				
Parcel 3	0.545	0.672	0.815	1.000			
Item 6	0.582	0.860	0.682	0.655	1.000		
Item 2	0.550	0.864	0.662	0.601	0.677	1.000	
Item 5	0.530	0.859	0.620	0.576	0.657	0.663	1.000

Note: * All correlations are significant at the p<.0001 level
Item 6=clear; Item 2=organized; Item 5=appropriate

Table 20

Correlations of the "expectations" construct

	Composite	Parcel 1	Parcel 2	Parcel 3	Item 2	Item 8
Composite	1.000					
Parcel 1	0.519	1.0000				
Parcel 2	0.452	0.514	1.000			
Parcel 3	0.507	0.626	0.609	1.000		
Item 2	0.492	0.784	0.529	0.576	1.000	
Item 8	0.381	0.781	0.337	0.458	0.323	1.000

Note: * All correlations are significant at the p<.0001 level
Item 2=clear; Item 8=demanding (but fair)

Table 21

Nine key questions on the recommended short form

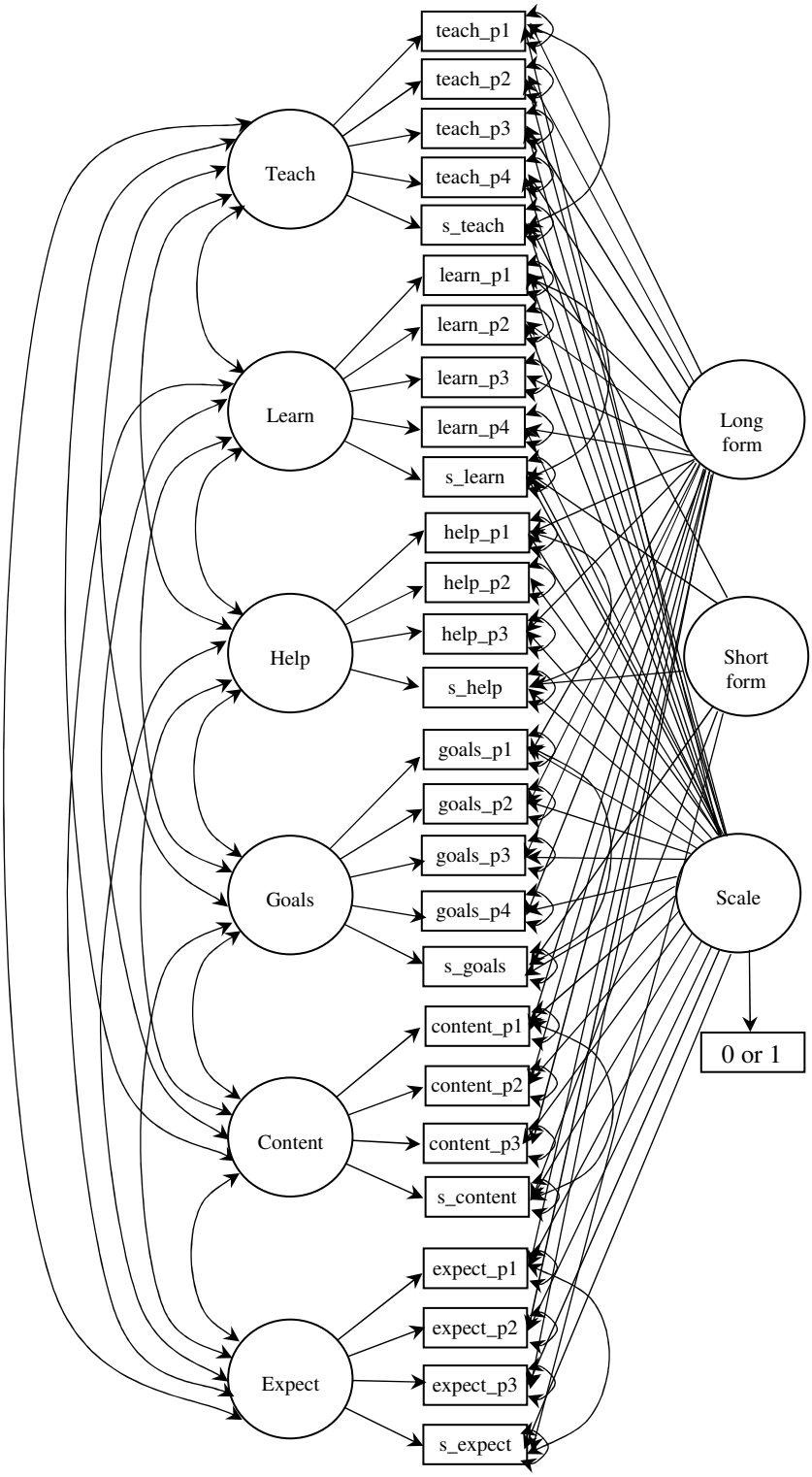
Kansas Board of Regents minimum requirements	Questions
Delivery of instruction	The instructor's teaching was clear, understandable, and engaging
Delivery of Instruction	The instructor was encouraging, supportive, and involved in my learning of the course material
Availability of Instructor	This instructor was available, responsible, and helpful
Delivery of Instruction	This instructor provided content and materials that were useful and organized
Whether goals and objectives were met	The instructor set and met clear goals and objectives for the course
Delivery of Instruction	What this instructor expected of me was well-defined and fair
Delivery of Instruction	What this instructor expected of me was appropriately challenging
Delivery of Instruction	The instructor demonstrated respect for me and my points of view
Assessment of learning	Compared with courses at a similar level, I would rate how much I learned as: much less, less, the same, more, much more

Table 22

"Happy medium" approach

The instructor's teaching was…	1	2	3	4	5
clear	◯	◯	◯	◯	◯
understandable	◯	◯	◯	◯	◯
detached	◯	◯	◯	◯	◯
unexciting	◯	◯	◯	◯	◯

Figure 1. Confirmatory factor analysis model



Appendix A

EVALUATION

Rate each item below of the following scale:

1=unsatisfactory; 2=below average; 3=average; 4=above average; 5=excellent

1. Has command of the subject.
2. Successfully communicates subject matter.
3. Is available to students on matter pertaining to the course.
4. Is sensitive to the response of the class.
5. Assigns readings, papers, projects, problems sets, etc., which are pertinent to the subject and helpful in learning it.
6. Provides meaningful critiques of students' efforts.
7. Is fair.
8. Overall, (s)he is an effective teacher.
9. The objectives of the course and the methods are clearly explained.
10. Overall, course goals and objectives are being achieved.
11. I would describe my learning in this class as.

REASON AND YEAR

Rate each item below on the following scale:

1=definitely false; 2=false; 3=neutral; 4=true; 5=definitely true

1. I took this course because I had little previous exposure to the area, and wanted to learn more about the subject.
2. I wanted to pursue a subject of previous interest and study.

3. The course was related to my career and professional interests.
4. I took this course to fulfill a major requirement.
5. I took this course to fulfill a school requirement.
6. I took this course largely because of the reputation of the course instructor.
7. I am a…

   freshman

   sophomore

   junior

   senior

   graduate student

   special student

COURSE IMPROVEMENT

These questions are to help your instructor assess particular strengths and weaknesses of this course. Your responses to this set will be used for course improvement rather than evaluation.

Rate each item below on the following scale:

D=disagree; MD=moderately disagree; N=neutral; MA=moderately agree; A=agree

1. The instructor's presentations were clear and understandable.
2. I felt free to express my ideas and questions in class.
3. Considering the nature of the course, the instructor was well prepared for each class session.

4. Appropriate attention was devoted to differing opinions and approaches to the subject matter.
5. I felt that the instructor was willing to help me outside of class.
6. Generally, I was prepared for each class session.
7. When questioned by class members, the instructor's responses were unclear and confusing.
8. Sufficient consideration was given to related fields and contemporary problems.
9. The instructor excessively dominated the class discussions.
10. This course aroused my intellectual curiosity.
11. The instructor seemed hostile toward students.
12. When there were discussions in class, I generally learned something from them.
13. I made an honest effort to learn in this course.
14. The instructor was dry and humorless.
15. The instructor's method of teaching this course needs revision.
16. The instructor raised questions and posed problems to the class.
17. When making generalizations, the instructor made good use of examples and illustrations.
18. The instructor made a genuine effort to get class members involved in the discussions.
19. The subject matter of this course seemed unimportant and insignificant to me.

20. The readings were too difficult.

21. The readings were appropriate in length.

22. I felt the exams stressed unwarranted memorization.

23. The instructor provided effective critiques of student projects.

24. The exam questions were phrased ambiguously.

25. The exams covered material emphasized in this course.

26. Overall, I felt that the instructor's grading was fair.

27. The requirements and deadlines of the course were made clear.

28. The various aspects of this course (lectures, discussions, readings, etc.) seemed to be integrated into a coherent whole.

Appendix B

Responses:

1=strongly disagree; 2=disagree; 3=neither agree nor disagree; 4=agree;

5=strongly agree

The instructor's teaching was:

1. clear
2. disorganized
3. understandable
4. engaging
5. simplistic
6. lacking energy
7. interesting
8. confusing
9. focused
10. thoughtful
11. detached
12. unenthusiastic
13. energetic
14. unexciting

Regarding my learning in this course, this instructor was:

1. indifferent
2. respectful

3. dismissive
4. unsupportive
5. encouraging
6. supportive
7. disinterested
8. helpful
9. uncaring
10. unconcerned
11. involved
12. careful to check my understanding of the course material

When I asked for help, this instructor was:

1. available
2. hard to contact
3. responsive
4. approachable
5. dismissive
6. unreceptive
7. helpful
8. uncaring
9. thorough
10. curt

Although I did not seek help, this instructor indicated that s/he would be:

1. available
2. hard to contact
3. responsive
4. approachable
5. dismissive
6. unreceptive
7. helpful
8. uncaring
9. thorough
10. curt

The learning goals and objectives of this course were:

1. clear
2. vague
3. appropriate
4. explicit
5. disjointed
6. ambiguous
7. demanding (but fair)
8. unchallenging
9. achieved/met
10. not accomplished

The materials and content of this course were:

1. ineffective
2. organized
3. unrelated
4. simplistic
5. appropriate
6. clear
7. off the topic
8. effective
9. not useful
10. helpful

The instructor's expectations of me were:

1. vague
2. clear
3. low
4. ambiguous
5. appropriate
6. explicit
7. unchallenging
8. demanding (but fair)

Appendix C

Responses:

1=strongly disagree; 2=disagree; 3=neither agree nor disagree; 4=agree;

5=strongly agree

1. The instructor's teaching was clear, understandable, and engaging.
2. This instructor was encouraging, supportive, and involved in my learning of the course material.
3. When I asked for help, this instructor was available, responsive, and helpful.
4. This instructor indicated that s/he would be available, responsive, and helpful.
5. This instructor provided content and materials that were clear, organized, and appropriate.
6. This instructor set and met goals and objectives for the course that were clear and appropriate.
7. This instructor's expectations of me were clear, demanding, and fair.
8. I am satisfied with my learning in this course.
9. Compared to other similar courses, I would rate how much I learned as:

   much below expected

   below expected

   expected/average

   above expected

   much above expected

Responses:

1=Not a reason; 2=somewhat important; 3=important; 4=very important

How important were the following reasons for taking this course?

1. Course fulfills a major or minor requirement.
2. Course fulfills an elective requirement.
3. Course fulfills a school requirement.
4. Course was not full (open).
5. Course was at a convenient time.
6. Course topic interests me.
7. Course instructor has a good reputation.

My student status is:

Undergraduate

Graduate

Other (non-degree, faculty, staff)

What year of study are you in?

1st

2nd

3rd

4th

5th

6th

7th or more

Did you complete readings/coursework?

Never

Rarely

Sometimes

Most of the time

Always

How many times per week did this class meet?

One

Two

Three

Four

Five

What grade do you expect in the class?

A

A-

B+

B

B-

C+

C

C-

D+

D

D-

F

Over the course of the semester, how many class meetings did you miss?